Main Figures and Table

Increasing metadata coverage of SRA BioSample entries using deep learning based Named Entity Recognition

Adam Klie1,2\*, Brian Y Tsui1,2\*, Shamim Mollah2,3,4, Dylan Skola1,2, Michelle Dow1,2, Chun-Nan Hsu1,5# and Hannah Carter1#

1Department of Medicine, Division of Medical Genetics, University of California, San Diego, La Jolla, CA 92093

2Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA 92093

3Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093

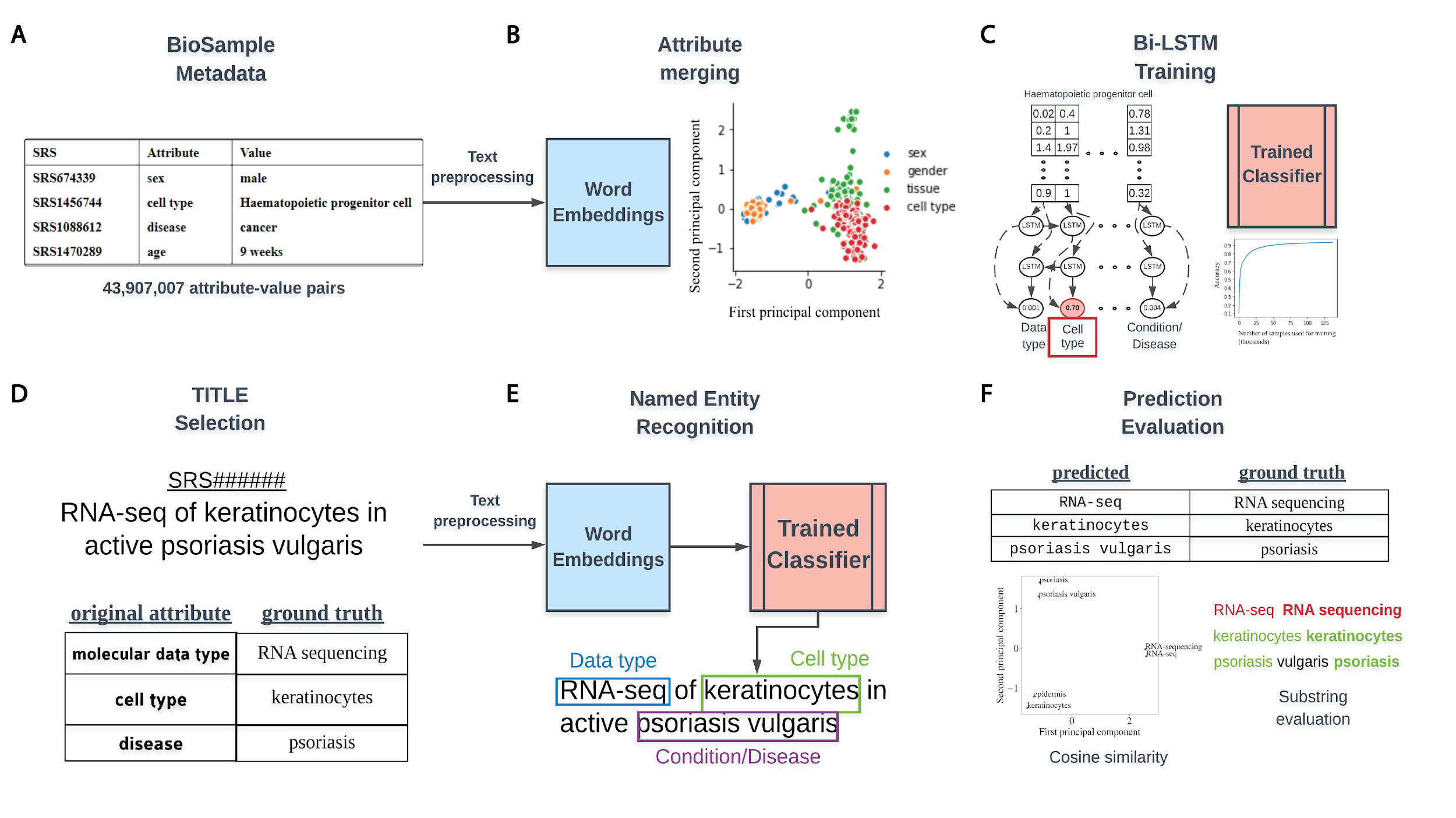
4Department of Genetics, Washington University in St. Louis, St. Louis, MO 63130

5Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093

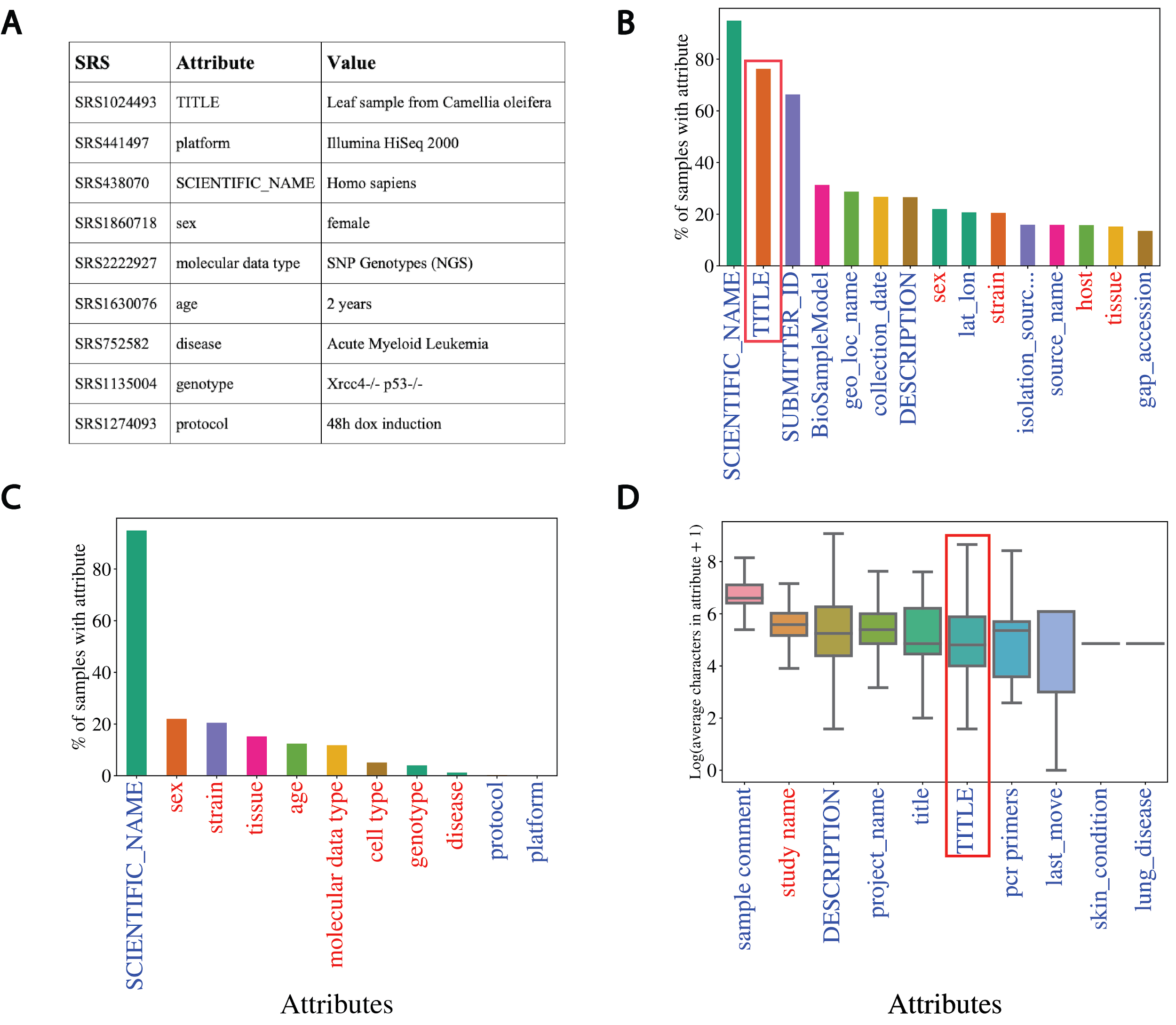
\*These authors contributed equally to this work

#To whom correspondence should be addressed

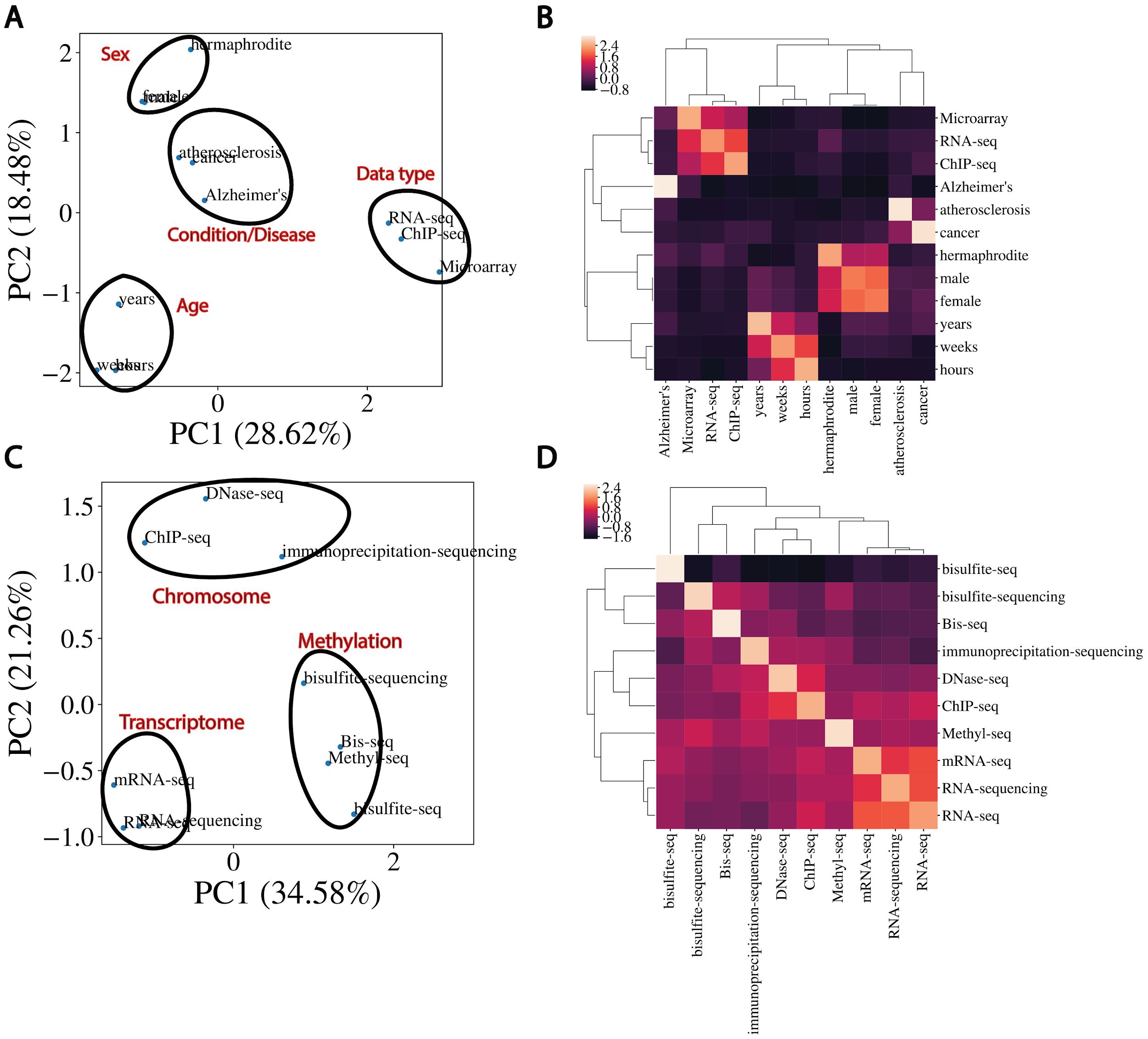
1. **Figure 1 -** Overview of classifier training and metadata prediction workflow.
2. **Figure 2 -** Missing metadata in SRA.
3. **Figure 3** - Word embeddings capture semantic similarities between words.
4. **Figure 4** - Performance of bi-LSTM in metadata category classification**.**
5. **Table 1** - Performance on the prediction of the 11 metadata categories from TITLEs.



**Figure 1.** Overview of classifier training and metadata prediction workflow. **(A)** A few examples of the 44 million attribute-value pairs in SRA BioSample. **(B)** Word embeddings of preprocessed values allowed for the clustering and merging of attributes that were similar in the embedding space. **(C)** A subset of attribute-value pairs was split into a training and test set and a bi-LSTM classifier was trained to identify 11 metadata categories. **(D)** TITLEs were selected as the free text for NER using the trained model. An example TITLE with associated ground truth labels is shown. **(E)** These TITLEs were preprocessed into n-grams and fed into the trained classifier after word embeddingto generate metadata predictions for the 11 categories. **(F)** Comparisons to ground truth metadata were done using substring matching and cosine similarity in the word embedding space.



**Figure 2.** Missing metadata in SRA. **(A)** Examples of SRA attribute-value pairs. Percentage of all samples that contained annotations for the **(B)** top 15 most used attributes and **(C)** the 11 selected attributes. X-axis shows attribute type and y-axis shows the percentage of total samples that used the given attribute. **(D)** Distributions of the average number of characters for the 10 longest (by mean) attributes in BioSample annotations of SRA. X-axis shows attribute type and y-axis shows the Log2(average characters) for a given attribute. Blue labels indicate a user defined attribute, red labels indicate a BioSample defined attribute. TITLE attribute in panels (B) and (D) is highlighted.



**Figure 3**.Word embeddings capture semantic similarities between words. **(A)** Dimensional reduction using PCA to visualize vectorized representation of a set of words encapsulating disease, age, data type and sex. **(B)** Corresponding correlation heatmap of the z-scored cosine similarity between words, with hierarchical clustering shown. **(C)** Same as (A) for data types. **(D)** Corresponding correlation heatmap of z-scored cosine similarity between words in (C).

## 

**Figure 4.** Performance of bi-LSTM in metadata category classification. **(A)** Accuracy, precision, recall, F1 score, and average AUROC calculated for all categories combined. **(B)** Accuracy of model classification on training set (y-axis) plotted against the number of training examples input (in thousands). **(C)** Percentage of each category correctly classified, shown as a heatmap, with predicted values on the x-axis and ground truth labels on the y-axis. **(D)** Receiver operating characteristic (ROC) curves for each category along with the average over all test set examples (micro average).

**Table 1.** Performance on the prediction of the 11 metadata categories from TITLEs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category Name** | **# TITLES** | **# Predicted** | **# Correct** | **Accuracy** |
| Age | 1000 | 224 | 106 | 47.32% |
| Cell type | 702 | 144 | 68 | 47.22% |
| Condition/Disease | 122 | 23 | 22 | 95.65% |
| Data type | 78 | 12 | 10 | 83.33% |
| Genotype | 595 | 130 | 81 | 62.31% |
| Platform | 275 | 30 | 11 | 36.67% |
| Sex | 190 | 4 | 4 | 100.00% |
| Species | 1000 | 855 | 811 | 94.85% |
| Strain | 1000 | 295 | 242 | 82.03% |
| Tissue | 1000 | 237 | 151 | 63.71% |
| Protocol | 18 | 0 | N/A | N/A |